The Advisability of using packages in data warehouse design

Manuel Serrano[1], Rafael Romero[2], Juan Carlos Trujillo[2], Mario Piattini[1]

[1]Alarcos Research Group
Department of Computer Science
E. S. Informática
University of Castilla - La Mancha
13071 Ciudad Real
{Manuel.Serrano, Mario.Piattini}@uclm.es

[2] Dept. de Lenguajes y Sistemas Informáticos
Universidad de Alicante
03080 Alicante
{romero, jtrujillo}@dlsi.ua.es

Abstract. Data warehouses are large data repositories integrating data from several sources that support decision making. Although, traditionally, data warehouses have been designed using the 'well-known' star schema, some design methodologies have come into existence in recent times. These new methodologies have not only focused on logical design: they also propose performing a conceptual modeling using UML. At present, it is widely accepted that modeling using packages simplifies the management and understanding of the designs.Until now, however, this statement has not been empirically proved in the data warehouse field. In this paper, we present an empirical study whose aim is to check whether using packages in designing data warehouses makes them more understandable.

Keywords. Data Warehouse, Design, UML.

1. Introduction

As transactional information systems have become more mature, agile and stable, business information needs have been changing at a similar pace. [24]. Nowadays, companies store large amounts of data, proceeding either from their operational systems, or bought at really low prices. One of the main problems that companies must face is that those data do not provide information on their own [9]. In order to solve that problem, companies are increasingly adopting products based on data warehouse technology. A data warehouse is defined as an integrated database that is used mainly in corporate decision making [14][19].

Data warehouses have become one of the most important trends in business computing, as they provide relevant and precise information for improving strategic decisions. In fact, some authors [15][22] have predicted a market of 12 to 15 billions dollars in the data warehouse field for the near future, and foresee annual increments of 20% [7].

Currently, several lifecycles and development techniques are being proposed for building data warehouses [1][6][12][16][17]. In these methodologies, one of the most important steps is modeling the data warehouse.

This modeling can be done at conceptual level – such as in the proposal of [4][10][11][6][20] -; at a logical level –for which the use of the star-schema design is universally accepted, providing good response times and an easy understanding of data and metadata from the points of view of the user and developer [17] -, and also at a physical level – as the designer has to choose the physical tables, the indices and data partitions which best represent the logical data warehouse thus facilitating its functionality. [3][15] -.

When producing a conceptual model of a data warehouse, object-oriented modeling languages such as UML are usually employed [21]. When modeling with UML, it is commonly accepted that using packages simplifies the design and improves the understanding of the schema. However, until now, that statement has not been proved in the data warehouse field empirically.

In this paper, we present an experiment we have carried out to investigate if conceptual models of data warehouse that use UML packages are more understandable than those that do not use packages.

The following section shows the hypotheses and goals of our experiment. In sections 3 and 4 we can see the planning and execution of that experiment. The fifth section deals with the analysis and interpretation of the study's results. In the final section we discuss the conclusions drawn from this paper.

2. The Goals of the Experiment

The goal of our empirical study is to determine if the UML conceptual models using packages are more understandable that those that do not use them.

Formally, we can define the goal of our experiment using GQM [26], as shown in table 1.

Table 1. Experiment goal definition

To analyse	The using of packages
To	Evaluate
With respect to	Understandability of UML models
From the point of view of	Designer
In the context of	Computer Science students

3. The Planning of the Experiment.

Our study is composed of one controlled experiment carried out by Ph.D. students from the University of Alicante (Spain) and one replica of that experiment which was developed by final-year Computer Science students from the University of Castilla – La Mancha (Spain). These experiments were performed using the recommendations that can be found in [8][18][23][27].

These experiments are part of a family which we expect to allow us to get the necessary knowledge to draw significant conclusions that can be applied in practice [2]. This family of experiments has been planned according to the method proposed by [8].

Although it could be advisable to use practitioners as experimental subjects, in reality it is quite difficult to have such subjects, so we did the experiments with students [5]. In this kind of experiments we consider that working with students, as we did, is perfectly valid, as they are the next generation entering the profession. Besides, the size of the actual difference between students and practitioners is small and the experimental tasks proposed in several experiments do not require industrial experience. Hence we can consider that it is viable to experiment with students, as is the case here... [2][13].

All the students taking part in the experiment have some knowledge in design and use of data warehouses, having studied these subjects as part of their academic training. The Ph.D. Students from the University of Alicante were those enrolled on an advanced course on data warehouses. Students from the University of Castilla – La Mancha were studying a subject in which data warehouses had been explained.

The proposal for the experiment was that it should be done by two different groups of subjects. The first group was to do a set of exercises using a complex data warehouse model and the second group had to work with a semantically equivalent model, but which had packages incorporated into the design. Models showed a real system representing a basic domain (supply chain) so as to be easily understood.

To do the experiment, we divided the subjects into two well-balanced groups. Each one of these would have to work on an exercise for one of the models (with packages or without packages). To form the groups we gave subjects a questionnaire analyzing their knowledge of UML, packages and data warehouses. Using the results of these questionnaires we divided the subjects in the balanced classification we have mentioned.

3.1. Hypotheses

The hypotheses of our experiment attempts to summarize the goals that we pursue in performing it.

H_0: There is no difference between the results obtained by the two experimental groups.

H_1: $\neg H_0$

3.2. Variables

The Independent variable in our experiment corresponds to the type of design used for creating the schemas (DESIGN_TYPE). Dependent variables are:

Effectiveness, defined as the number of correct answers with respect to the number of total questions.

$$Effectiveness = \frac{\text{Number of Correct Answers}}{\text{Number of Questions}}$$

Efficiency, which can be calculated as the number of correct answers per time unit.

$$Eficiency = \frac{\text{Number of Correct Answers}}{\text{Time}}$$

Our goal is to observe how the design type affects these two dependent variables, and to see if there is any kind of differences in the behaviour of the subjects, depending on which exercise they were doing.

3.3. Objects used in the experiment

In this experiment we used a conceptual data warehouse schema, which was designed using packages, and another schema, semantically equivalent to the first one, designed without using packages. Both schemas represented an UML design of a supply chain data warehouse.

Subjects were given the experimental documentation in which they found the instructions for carrying out the exercises and an example of how to do the proposed task- then they could find the design of the data warehouse. After that, a set of exercises were proposed. Subjects were responsible for writing down their starting and finishing time for the exercises.

4. The Execution of the Experiment.

The experiments were carried out in two sessions: in one of them the subjects did the knowledge-assessment test. After analysing these tests, we classified the subjects into two well-balanced groups. In the second session, we did the experimental tasks.

At the beginning of the second session, we gave an intensive explanation about what kind of exercises the subjects were going to find, how they had to deal with them and what kind of documentation they would receive in order to do the experiment. However, subjects were not aware of the aspects we aimed to study, neither had they any knowledge regarding the hypotheses stated above. In addition, they did not know that there were two kinds of schemas, nor to which one they were assigned.

A supervisor controlled the experiment, ensuring that .it was being carried out correctly. This person answered the questions that arose during the course of the experiment.

4.1. Validity of results

To avoid some possible threats to the results of the experiment, we decided to take some measures to improve the whole empirical work:

- Experimental subjects were divided into two well-balanced groups according to their knowledge of UML, packages design and data warehouses.
- The domain of the schemas was common and simple enough, so there were no problems in understanding them on this count.
- This was the first time that the subjects had taken part in an experiment like this. Persistence effects were lessened accordingly.
- The Subjects were motivated, as the exercises they did in the experiment were part of the knowledge they are supposed to acquire in the subject. Also, as they were final-year students and Ph.D. students, they were interested in learning how to conduct an empirical study like this.
- Subjects were not allowed to talk to each other. We did not permit them to look at their companions' exercises either.
- All their questions and doubts were answered by the supervisor conducting the experiment.
- In order to avoid problems with time recording, we projected a digital clock onto the wall during the experiment.

5. Analysis and Interpretation

Before analysing the results of the experiment, we studied the collected data with the aim of erasing the values given by those subjects whose behaviour was not comparable to the rest of the subjects of the sample. Firstly, we noticed that subject number 13 of the replication experiment had neither answered the exercises correctly nor had written down the time spent in doing them. This subject was not considered in the rest of the analysis.

As a second step we analysed the data, looking for outliers. To perform this kind of analysis we made box plots for the dependent variables of our study. These box plots are shown in figures 1 and 2 (original experiment) and 3 and 4 (replica).

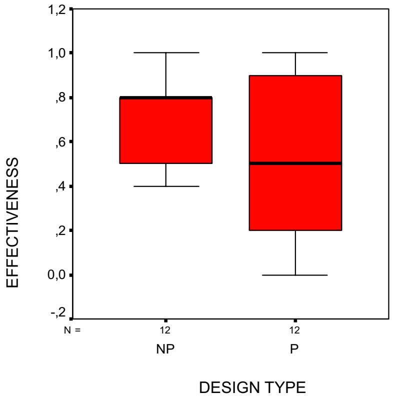

Fig. 1. Box plots for effectiveness variable in the original experiment

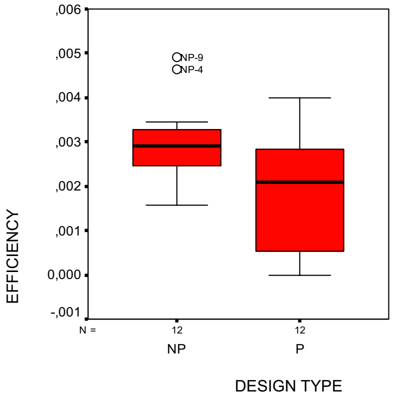

Fig. 2. Box plots for efficiency variable in the original experiment.

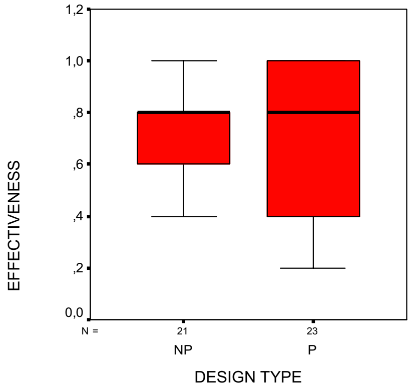

Fig. 3. Box plots for effectiveness variable in the replica experiment

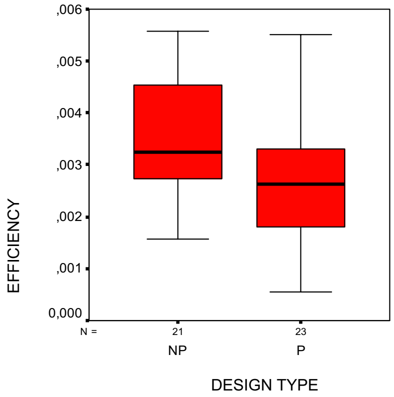

Fig. 4. Box plots for the efficiency variable in the replica experiment

As we can see in the box plots (figures 1, 2, 3 and 4) there are neither strange nor extreme values, except when we are considering the efficiency of the original experiment. In that experiment subjects 4 and 9 have strange values and were not considered in the analysis stage. The rest of the data were used in the statistical analysis.

Before doing the statistical analysis we established a significant level of α = 0.1, as one way to be able to increase the power of the statistical tests (that is, the probability of rejecting our hypotheses when they are false) is increasing the significance level.

Bearing in mind the study goals, the experiment configuration and the collected data we used a multivariate ANOVA test [25].

The results that we obtained from the application of the statistical test to the data collected in the original experiment are shown in table 2. Analysing the significance value, with respect to the α level (α = 0,1), we can see that the significance value for the effectiveness variable is greater than α, and thus we cannot reject the null hypothesis (H_0: There is no difference due to the design type between the effectiveness in answering the exercises). On the other hand, we can see that the significance values for the efficiency variable are less than α and we can conclude that there is difference between the results of the subjects who used the design with or without packages, respectively.

In order to see which of the two design methods has the best efficiency we can observe in table 4 that the average efficiency of the subjects working with the schema with packages was less than the average efficiency of the other group.

By observing tables 3 and 4 we can see that in the replica experiment, the results are the same.

Table 2. ANOVA Analysis of the original experiment

ANOVA								
Source	Dependent variable	Type III Square Sum	DF	Quadratic Mean	F	Significance	Parameter of non-centrality	Observed Power
DESIGN TYPE	EFFECTIVENESS	0,166666667	1	0,16666667	1,76848875	0,19719336	1,768488746	0,3626088
	EFFICIENY	3,886E-06	1	3,886E-06	3,37893039	0,08093526	3,37893039	0,55229029

Table 3. ANOVA Analysis of the replica experiment

ANOVA								
Source	Dependent variable	Type III Square Sum	DF	Quadratic Mean	F	Significance	Parameter of non-centrality	Observed Power
DESIGN TYPE	EFFECTIVENESS	0,02620742	1	0,02620742	0,48369936	0,49058189	0,48369936	0,17831571
	EFFICIENY	7,6837E-06	1	7,6837E-06	5,67131025	0,02185308	5,67131025	0,75742899

Table 4. Mean differences in both experiments

Efficiency Means	Original Experiment	Replica
Packages	0,0018	0,0027
No-Packages	0,0027	0,0035

As a conclusion from the analysis of the experiment, we can state that there are no differences in terms of effectiveness when working with conceptual schemas of data warehouses, regardless of the design method (with or without packages) we use. However, with respect to efficiency, we can find that there are differences, because the time spent in handling those schemas designed using packages is greater than in handling those without packages, and this issue lessen its efficiency.

In spite of these results being opposed to common belief (that is better to design using packages), we must be aware that the result can be influenced by the fact that the experiments were carried out using only pen and paper, without the help of a tool for navigating through the schema. In this way, the schema designed by not using packages were represented on only one sheet, while subjects using the other design had to go through several sheets in order to find what they were looking for, spending more time in the exercises. Also, the size of the schema may perhaps affect the results. We must go on doing more replicas of this experiment, to obtain more conclusive results; but, at this point in time, we can conclude that it is not always more efficient to design conceptual data warehouse schemas using packages.

6. Conclusions

Data warehouses are one of the main industrial trends in information systems, as they help in strategic decision making.

Most of the proposed data warehouse design methodologies are based on making designs of the data warehouse at different levels, conceptual, logical and physical.

Focusing on data warehouse conceptual modeling, it is the normal course to use object-oriented designs and UML modeling language to design this kind of conceptual schemas.

When using UML, it is widely accepted that using packages organizes the schema and simplifies its use. Although this statement is commonly held to be true, nobody has empirically demonstrated its veracity in the data warehouse field. We have done an experiment to prove whether or not this is the case...

Our experiment concludes that there is not likely to be any difference regarding effectiveness between subjects working with the schemas designed with or without packages. With respect to efficiency, it seems that subjects working with schemas without packages are more efficient than those using the same schema designed with packages.

The way in which the experiment took place can affect the result, because the exercises were done using only pen and paper. Our next goal is to prove if these results are valid when working with a tool that helps us in navigating through the schemas.

Acknowledgement

This paper is part of the CALIPO research project, partially funded by the Dirección General de Investigación del Ministerio de Ciencia y Tecnología (TIC2003-07804-C05-03). It is also part of the research Project, partially funded by the Conserjería de Ciencia y Tecnología de la Junta de Comunidades de Castilla-La Mancha (PCC-03-003-1).

We also want to kindly thank the participating students from the Universities of Alicante and Castilla – La Mancha (Spain) for taking part in the proposed experiments and in helping us in this research.

References

[1] Anaya, V., Pérez, J. y Celma, M. Diseño de Almacenes de Datos: Sentido y Simplicidad. in VIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2003). 2003. Alicante.

[2] Basili, V., Shull, F. y Lanubile, F., *Building knowledge through families of experiments.* IEEE Transactions on Software Engineering, 1999. **25(4)**: p. 435-437.

[3] Bouzeghoub, M. y Kedad, Z., *Quality in Data Warehousing*, in *Information and database quality*. 2002, Kluwer Academic Publishers.

[4] Cabbibo, L. y Torlone, R. *A logical approach to multidimensional databases*. in *Sixth Internationational Conference on Extending Database Technology (EDBT'98)*. 1998. Valencia, Spain: Lecture Notes in Computer Science 1377, Springer-Verlag.

[5] Carver, J., Jaccheri, L., Morasca, S. y Shull, F. *Issues in Using Students in Empirical Studies in Software Engineering Education*. in *9th International Software Metrics Symposium (METRICS'03)*. 2003.

[6] Cavero, J., Marcos, E. y Piattini, M. *Metodología para el Diseño de Almacenes de Datos: Etapa de Modelado Conceptual*. in *4º Encontro para a Qualidade nas Tecnologias de Informação e Comunicações (QUATIC 2001)*. 2001. Lisbon, Portugal.

[7] Chenoweth, T., Schuff, D. y St. Louis, R., *A method for developing Dimensional data marts.* Communications of the ACM, 2003. **46**(12): p. 93-98.

[8] Ciolkowski, M., Shull, F. y Biffl, S. *A Family of Experiments to Investigate the Influence of Context on the Effect of Inspection Techniques*. in *6th International Conference on Empirical Assessment in Software Engineering (EASE)*. 2002. Keele (UK).

[9] Gardner, S. R., *Building the data warehouse.* Communications of the ACM, 1988. **41**(9): p. 52-60.

[10] Golfarelli, M., Maio, D. y Rizzi, S. *Conceptual design of data warehouses from E/R schemes*. in *31st Hawaii International Conference on System Sciences*. 1998. Hawaii, USA.

[11] Golfarelli, M. y Rizzi, S., *Designing The Data Warehouse: Key Steps and Crucial Issues.* Journal of Computer Science and Information Management, 1999. **2**(3).

[12] Hammergren, T., *Data Warehousing Building the Corporate Knowledge Base*. 1996, Milford: International Thomson Computer Press.

[13] Höst, M., Regnell, B. y Wholin, C. *Using Students as Subjects - A comparative Study of Students & Professionals in Lead-Time Impact Assessment*. in *4th Conference on Empirical Assessment & Evaluation in Software Engineering (EASE)*. 2000. Keele University, UK.

[14] Inmon, W. H., *Building the Data Warehouse*. Third Edition ed. 2002, USA: John Wiley and Sons.

[15] Jarke, M., Lenzerini, M., Vassiliou, Y. y Vassiliadis, P., *Fundamentals of Data Warehouses*. second edition ed. 2002: Springer-Verlag.

[16] Kelly, S., *Data Warehousing in Action*. 1997: John Wiley & Sons.

[17] Kimball, R., Reeves, L., Ross, M. y Thornthwaite, W., *The Data Warehouse Lifecycle Toolkit*, J.W.a. Sons, Editor. 1998: USA.

[18] Kitchenham, B., Pfleeger, S., Pickard, L., Jones, P., Hoaglin, D., El Emam, K. y Rosenberg, J., *Preliminary Guidelines for Empirical Research in Software Engineering.* IEEE Transactions on Software Engineering, 2002. **28**(8): p. 721-734.

[19] Lechtenbörger, J. *Data Warehouse Schema Design*. in *Datenbanksysteme für Business, Technologie und Web (BTW 2003)*. 2003. Leipzig (Germany.

[20] Luján-Mora, S., Trujillo, J. y Song, I.-Y. *Extending UML for Multidimensional Modeling*. in *5th International Conference on the Unified Modeling Language (UML 2002)*. 2002: LNCS 2460.

[21] OMG, *OMG Unified Modeling Language Specification; versión 2.0.* 2005, Object Management Group.

[22] Parkes, C., *Data Warehousing: The economy isn't the only reason the data warehouse industry is stumbling*. 2002: Enterprise Systems.

[23] Perry, D., Porte, A. y Votta, L., *Empirical Studies of Software Engineering: A Roadmap.* Future of Software Engineering, Ed. Anthony Finkelstein, ACM, 2000: p. 345-355.

[24] Phipps, C. y Davis, K. C. *Automating Data Warehouse Conceptual Schema Design and Evaluation*. in *4th International Workshop on Design and Management of Data Warehouses (DMDW 2002)*. 2002. Toronto (Canada).

[25] SPSS, *SPSS. Syntax Reference Guide. SPSS Inc.* 2001.

[26] Van Solingen, R. y Berghout, E., *The Goal/Question/Metric Method: A practical guide for quality improvement of software development*. 1999: McGraw-Hill.

[27] Wohlin, C., Runeson, P., Höst, M., Ohlson, M., Regnell, B. y Wesslén, A., *Experimentation in Software Engineering: An Introduction.* 2000: Kluwer Academic Publishers.